DOI:
10.1109/TITS.2015.2505304

Document Version
Peer reviewed version

IEEE Transactions on Intelligent Transportation Systems

Efficient cellular load balancing through mobility-enriched vehicular communications

Journal:	*Transactions on Intelligent Transportation Systems and Intelligent Transportation Systems Magazine*
Manuscript ID	T-ITS-14-11-0712.R2
Manuscript Type:	Regular Papers
Date Submitted by the Author:	n/a
Complete List of Authors:	Kolios, Panayiotis; King's College London, Department of Informatics, Centre for Telecommunication Research Papadaki, Katerina; London School of Economics, Department of Management Friderikos, Vasilis; King's College London, Department of Informatics, Centre for Telecommunication Research
Keywords:	wireless communication, Vehicles, Energy management, Optimization methods, cellular networks
Abstract:	Supporting effective load balancing is paramount for increasing network utilization efficiency and improving the perceivable user experience in emerging and future cellular networks. At the same time, it is becoming increasingly alarming that current communication practices lead to excessive energy wastes both at the infrastructure side and at the terminals. To address both these issues, this paper discusses an innovative communication approach enabled by the implementation of device-to-device (d2d) communication over cellular networks. The technique capitalizes on the delay tolerance of a significant portion of Internet applications and the inherent mobility of the nodes to achieve significant performance gains. For delay tolerant messages, a mobile node can postpone message transmission $-$ in a store-carry and forward manner $-$ for a later time to allow the terminal to achieve communication over a shorter range or to postpone communication to when the terminal enters a "cooler cell", before engaging in communication. Based on this framework, a theoretical model is introduced to study the generalized multihop d2d forwarding scheme where mobile nodes are allowed to buffer messages and carry them while in transit. Thus, a multi-objective optimization problem is introduced where both the communication cost and varying load levels of multiple cells are to be minimized. We show that the mathematical programming model that arises can be solved efficiently in time. Further, extensive numerical investigations reveal that the proposed scheme is an effective approach for both energy efficient communication as well as offering significant gains in terms of load balancing in multicell topologies.

Efficient cellular load balancing through mobility-enriched vehicular communications

Panayiotis Kolios*, Katerina Papadaki†, Vasilis Friderikos‡

***Abstract*—Supporting effective load balancing is paramount for increasing network utilization efficiency and improving the perceivable user experience in emerging and future cellular networks. At the same time, it is becoming increasingly alarming that current communication practices lead to excessive energy wastes both at the infrastructure side and at the terminals. To address both these issues, this paper discusses an innovative communication approach enabled by the implementation of device-to-device (d2d) communication over cellular networks. The technique capitalizes on the delay tolerance of a significant portion of Internet applications and the inherent mobility of the nodes to achieve significant performance gains. For delay tolerant messages, a mobile node can postpone message transmission – in a store-carry and forward manner – for a later time to allow the terminal to achieve communication over a shorter range or to postpone communication to when the terminal enters a cooler cell, before engaging in communication. Based on this framework, a theoretical model is introduced to study the generalized multihop d2d forwarding scheme where mobile nodes are allowed to buffer messages and carry them while in transit. Thus, a multi-objective optimization problem is introduced where both the communication cost and varying load levels of multiple cells are to be minimized. We show that the mathematical programming model that arises can be solved efficiently in time. Further, extensive numerical investigations reveal that the proposed scheme is an effective approach for both energy efficient communication as well as offering significant gains in terms of load balancing in multicell topologies.**

***Index Terms*—Load balancing, Energy efficiency, Store carry and forward relaying, Device to device communication, Wireless routing, Cellular networks.**

I. Introduction

With the increasingly high availability of rich content Internet applications on mobile devices and more recently vehicular terminals (which are driven by user demand), data use surging is expected to increase, severely compromising system performance and eventually (as a by-product) overall user experience. Residing on the supporting neighboring cells of hot-spot areas to alleviate the problem, numerous techniques are proposed in the literature to either decide on the initial user association policies [1]-[4], cell reselection algorithms [5]-[8] or network-driven handover procedures [9]. However, the majority of these techniques has as a salient assumption real-time constraints on the requested traffic and thus decisions are made on the instantaneous cell loading conditions. In [10] batch processing of data requests is considered; illustrating the potential load balancing improvements that can be achieved by collectively considering the load assignment problem over the available infrastructure nodes. In contrast to the latter work, hereafter we labour a broader range of message delivery delays where elastic services can handle delays in the order of few seconds with no degradation in the perceivable user experience. In that respect, a mobile node (e.g. vehicular terminal) could potentially postpone communication for a later time instance when it enters the serving area of a neighboring cell, in effect reducing the load imbalance across the network. Further, by postponing communication a mobile node can approach closer to the serving infrastructure node before initiating the transmission and hence reducing the physical (Euclidean) communication distance that compromises communication. Therefore, both the inter-cell load levels and intra-cell energy consumption can be reduced by exploiting the available message delivery delays and exploring the feasible forwarding paths that can be formed by the mobile nodes.

A. Background and Related Work

Node mobility has been previously considered as a key element to achieve communication in intermittently connected networks (i.e., Delay Tolerant Netwoks)[11][12], to increase the capacity of ad-hoc and cellular networks ([13] and [14] respectively) and to improve the energy efficiency by reducing the physical communication distance between communicating nodes [15]-[17]. In this work we investigate an additional benefit branching as a result of mobility and delay tolerance; that of energy efficient load balancing in cellular systems. Notably, by postponing communication, a mobile node can carry information while in transit and opt for a better serving cell with favourable load conditions. Hence, communication can be achieved via a store-carry and forwarding (SCF) paradigm under the deadline constraints imposed by the initiated service. Note that the SCF scheme considered hereafter differs drastically from the ones previously proposed for delay tolerant networks (DTNs). Firstly in DTNs, and due to the absence of end-to-end connectivity, information is stored by nodes and opportunistically forwarded at node encounters. Here, due to the presence of the infrastructure network, all nodes are able

*Department of Informatics, Centre for Telecommunications Research, King's College London, Strand, WC2R 2LS, London, England, e-mail:*panayiotis.kolios, ‡vasilis.friderikos@kcl.ac.uk

†Department of Management, London School of Economics, Houghton Street, London WC2A 2AE, London, England, e-mail: k.p.papadaki@lse.ac.uk

Manuscript received ...

to communicate directly with at least one base station (BS) and thus timely deliveries can be guaranteed. Secondly, in DTNs due to the unpredictable mobility patterns, messages are replicated at node encounters (in a broadcast or more intelligent manner) to increase the probability for successful communication. For the applications we consider here, informed routing decisions are made by the infrastructure units that have instantaneous knowledge regarding all node positions within the coverage area of the BSs. The proposed load balancing scheme will be further propelled as device to device (d2d) communication will be a supported feature in LTE Release 12 and beyond.

B. Paper contributions

The multihop scenario is considered where mobile nodes, in addition to sources of data, act as relays for other nodes in the cell that are either fixed or mobile. We therefore consider the existence of the following three entities in the network: a) user terminals (UTs), b) vehicular relays (VRs) and c) base stations (BSs). Illustratively, figure 1 shows an arbitrary cellular network topology. In a simple instantiation of the problem, a VR (which can also be a source node) buffers information messages while in transit and communication is achieved via a single hop store-carry and forward path (link 5 in fig. 1). In addition, a VR can forward data to another VR (link 3 in fig. 1) which in turn can transmit to a VR ahead or to the BS (link 4 in fig. 1) in a multihop fashion. Further, a UT can transmit directly to the BS (link 1 in fig. 1) or employ a VR for message forwarding (link 2 in fig. 1). Note that this scheme is also applicable for the downlink of information from the BS to the UTs/VRs in the network. For example, a BS can postpone communication until a UT enters its serving cell or transmit to a VR that in turn buffers the data and carries it towards the destination UT found in the serving area of another cell.

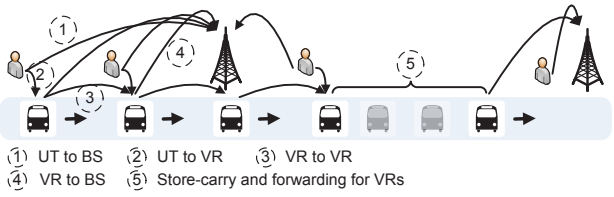

Fig. 1: This figure illustrates the proposed network layout where VR nodes can buffer information while in transit via store-carry and forwarding. The uplink case is considered in this figure however the scheme is applicable for the downlink case as well.

A prominent application of d2d and an early adopter will be vehicular to vehicular (v2v) and vehicular to infrastructure (v2i) communications. Advanced optimized transmissions that allow for efficient communication (and effectively lower latencies as well) can be deemed as of paramount importance for the support of innovative new services. As previously argued, the actual inherent mobility in v2v and v2i communications is utilized as an additional resource to optimize the overall system performance. While load balancing provides obvious benefits for all cellular users, energy efficiency for vehicular terminals is also critical. The enormous volume of sensor data from automated driving cars that needs to be communicated to the infrastructure (current BMW cars produce a staggering 10Gps of such sensor data [18]) creates significant challenges for energy consumption and could dramatically impact the battery capacity especially for electric vehicles that already suffer from range anxiety.

In addition, the proposed approach can be envisioned as a natural fit for the emerging so called C/U split plane wireless architectures, where macro cells act as the network control plane (C-plane) that efficiently manage in real time resources of a large number of high capacity small cells (including future mmWave-based cells, aka 5G) that can serve mobile users ”on demand” (U-plane) [28]. The concept of C/U split has been envisioned in the setting of small so-called phantom cells that will serve mobile users (U-plane) while being controlled by a macro base station operating at different (lower) frequency bands [29], [30]. In that setting the proposed technique can be utilized by the macro-controller (macro-cell) to decide allocation of users to different high capacity small cells.

The contributions of this paper are as follows:

- Under the assumptions of delay tolerant traffic we have proposed a novel multi-objective mathematical model for calculating data traffic routes, where users can use traditional relaying but also store carry and forward relaying, in order to handle load balancing within the cells while at the same time minimizing energy consumption and data traffic delay.
- Extensive simulations have been performed that show that the store carry and forward routes are frequently used along with traditional relaying routes providing significant gains in terms of energy efficiency. The trade-offs between load balancing, energy consumption and delay have been thoroughly studied.
- Our model was also tested on a realistic network topology on road segments close to Kings College London campus (in central London) with two base stations and a single cellular operator, where mobility traces were generated for real traffic thus showing that our model can be used in real traffic environment.
- Our model is in-line with emerging 5G wireless network architectures where store-carry and forward decisions will be taken by the macro-controller managing a large number of high capacity small cells.

The paper is structured as follows. In Section II the network setup is explained and in Section III an optimization problem is derived for minimizing both the load imbalance in the network and the communication cost of the elected forwarding paths. Section IV considers a numerical investigation on the performance of the proposed scheme while Section IV-C provides detailed numerical investigations for different network setups. Finally, Section VI concludes the paper.

II. System Model

In this section we first consider the network structure under investigation and define the communication model and load balancing models. Table I contains variable definitions to be used as a reference for the derivations that follow.

We consider the uplink case of cellular operation, however the model is applicable to the downlink case as explained above. A constellation of $\mathcal{C} = \{1, \ldots, C\}$ cells is considered each with cell radius of R meters. For illustration reasons we initially assume a 1-dimensional realization of the network topology, and more general cases are considered in section IV-C. Further, $\mathcal{M} = \{1, \ldots, M\}$ uniformly distributed users are considered to be active while $\mathcal{N} = \{1, \ldots, N\}$ relay nodes travel along each direction of a bi-directional stretch of road. To differentiate between source and relay nodes we assume that all users (source nodes) are static or slow moving and thus their position does not change abruptly. Note that this assumption does not restrict the model in any way, any node can be considered as a source node in general and the model is still applicable. Finally, with v_j we denote the velocity of mobile node $j \in \mathcal{N}$.

A. Communications Model

For the communication model we assume that all nodes can transmit and receive from a single other node at any time instance. All users have a single message of F (bits) to communicate to the BS (the case of variable size messages is considered in appendix B) and all links are able to transmit at a data rate of B (bps). We further consider three sources of energy consumption: 1) the circuit energy consumption at the transmitter side, e_t, 2) the energy consumed to receive a message at the receiver, e_r and 3) the energy consumed by the power amplifier at the transmitter side, e_d. Large scale propagation losses are considered with signal attenuations following a two-stage model. For transmission distances less than a threshold distance d_{brake}, the free space model is considered, while for large distances the plane-earth model is employed.

The physical distance between communicating nodes may change during transmission as some nodes are mobile. We define $\bar{v}_{ij}$ to be the relative velocity between communicating nodes i, j. Also D_{ij} is the initial distance between the transceiver pair i, j. With $g(D_{ij}, t)$ we denote the absolute distance at time t between the transceiver pair i, j during transmission and is defined as follows:

$$g(D_{ij}, t) = |D_{ij} - \bar{v}_{ij} t|. \tag{1}$$

Based on the above equation, the total energy consumption between nodes i and j can be computed as follows:

$$f_{ij} = (e_r + e_t) B\tau + B e_d \int_0^{\tau} g(D_{ij}, t)^{\eta} dt, \tag{2}$$

where $\tau = F/B$ is the communication time and η expresses the pathloss exponent.

B. Load Balancing Model

Balancing the load across the C BSs is equivalent to minimizing the variance of these loads. We let S_k be the initial load (for example in bps) for each $k \in \mathcal{C}$, q be the increase in rate of accepting each user request, and U_k be the maximum capacity in terms of rate at BS k. We define integer variables y_k to be the total number of user requests accepted by BS k, and let binary variables z_k^m take the value 1 only if user m is satisfied by BS k.

Using the above notation, the variance between the loads at the different BSs after accepting the requests is as follows:

$$\text{Var} = \sum_{k=1}^{C} \left[y_k q + S_k - \frac{\sum_{j=1}^{C} (y_j q + S_j)}{C} \right]^2 \tag{3}$$

where the new load at BS j is $y_j q + S_j$. To balance the load between BSs, we formulate the following optimisation problem:

$$\text{(P1)} \min \ \text{Var} \tag{4}$$

$$\text{s.t.} \sum_{k=1}^{C} z_k^m = 1 \ \forall\, m \in \mathcal{M} \tag{5}$$

$$\sum_{m=1}^{M} z_k^m = y_k \ \forall\, k \in \mathcal{C} \tag{6}$$

$$y_k q + S_k \leq U_k \ \forall\, k \in \mathcal{C} \tag{7}$$

$$z_k^m \in \{0, 1\}, y_k \geq 0, y_k \in \mathbb{Z}. \tag{8}$$

Constraints (5) ensure that all user requests are accepted (serviced) by the system, whereas constraints (6) define the number of requests accepted by BS k. Finally, constraints (7) set an upper bound on the total load of each BS. In effect, the actual upper bound of variables y_k is u_k:

$$u_k = \left\lfloor \frac{U_k - S_k}{q} \right\rfloor \tag{9}$$

In the sequel, an alternative formulation of the load balancing problem using network flows is derived and we show that it is equivalent to problem (P1). Consider the network on figure 2. We define a minimum cost flow problem on this network, where source node $m \in \mathcal{M}$ has supply $+1$ and sink node K has demand M. We let

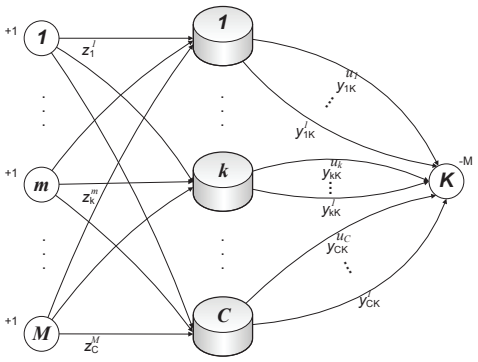

Fig. 2: Network flow model for load balancing.

TABLE I: Parameter and Variable Definitions

Notation	Definition	Notation	Definition
$\mathcal{C}$	set of BSs indexed 1 to C	U_k	max capacity in terms of rate at BS k
$\mathcal{M}$	set of static users indexed 1 to M	q	Minimum data traffic per request
$\mathcal{N}$	set of relay nodes indexed 1 to N	y_k	Volume of user requests accepted at k
F	Message size (bits)	z_k^m	Flow variable for user m on cell k
B	Data rate (bps)	y_{kK}^n	Flow variable for cell k to sink K for request n
e_t	Circuit energy consumption at the transmitter	u_k	upper bound of variables y_k
e_r	Circuit energy consumption at the receiver	x_{ij}	Flow variable between nodes i and j
e_d	Energy consumption of the transmitterś power amplifier	$\mathcal{T}$	Time horizon
d_{brake}	Threshold distance for change in the propagation losses model	G	$= G(V, L)$; graph on space-time network
v_j	Velocity of node j	V_p	set of (i, a), where $i \in \mathcal{N}$ at time epoch a
$\bar{v}_{ij}$	Relative velocity between nodes i and j	BSC	super sink node for graph G on space-time network
D_{ij}	Distance between nodes i and j	V	$= \{\mathcal{M}, V_p, \mathcal{C}, BSC\}$
τ	$= F/B$; Communication duration	$b(i)$	supply and demand on nodes in G
η	Pathloss exponent	λ	weighting parameter between energy and delay
S_k	Data traffic load of cell k	γ	weighting parameter between load balancing and communication cost
f_{ij}	total energy consumption between nodes i and j		

variables z_k^m denote the flow of the arc from each source node m to node k and we assume that they have capacity one and cost zero. Since the available capacity of each BS k is u_k, we introduce u_k arcs of capacity one from BS k to sink node K. We denote the flow of these arcs by y_{kK}^n where $k = 1, \ldots, C$ and $n = 1, \ldots, u_k$. The cost on link y_{kK}^n defined in equation (10) below reflects the increase in utilization of accepting the n^{th} request over BS k and thus provides a cost incentive for load balancing.

$$w_k(n) = q\, S_k + n\, q^2 \tag{10}$$

Based on the above definitions, the minimum cost flow formulation of the above network is:

$$(\text{P2}) \min \sum_{k=1}^{C} \sum_{n=1}^{u_k} w_k(n) y_{kK}^n \tag{11}$$

$$\text{s.t.} \sum_{k=1}^{C} z_k^m = 1, \ \ \forall m \in \mathcal{M} \tag{12}$$

$$\sum_{m=1}^{M} z_k^m - \left(\sum_{n=1}^{u_k} y_{kK}^n \right) = 0, \ \ \forall k \in \mathcal{C} \tag{13}$$

$$- \sum_{k=1}^{C} \left(\sum_{n=1}^{u_k} y_{kK}^n \right) = -M \tag{14}$$

$$z_k^m \in \{0, 1\}, \forall k \in \mathcal{C}, \forall m \in \mathcal{M} \tag{15}$$

$$y_{kK}^n \in \{0, 1\}, \forall k \in \mathcal{C}, \forall n = 1, \ldots, u_k \tag{16}$$

Constraints (12), (13) and (14) are the flow conservation constraints for the source nodes $m \in \mathcal{M}$, basestations $k = 1, \ldots, C$ and the sink node K, respectively. The capacity constraints for the flow variables are given by (15) and (16) and since each request is services by a single basestation, the integrality of these flows is a natural assumption. However, the integrality constraints can be relaxed while guaranteing integer optimal solutions for problem (P2).

Proposition 1: The linear programming relaxation of problem (P2) guarantees to give integer optimal solutions.

Proof: Min-cost flow formulations with integer supply/demand and link capacities (as is the case here), have integer optimal solutions according to the integrality principle [19]. ■

Furthermore, we have:

Proposition 2: Problem (P1) is equivalent to problem (P2).

The proof can be found in appendix A. Importantly, given that the two problems are equivalent, then the load balancing problem can be solved using efficient linear programming algorithms,[19]. Noticeably, the derivations above have been based on the fact that each request imposes an equal increase in utilization in the system. Problem (P2) can be extended to include the alternative case where each user request imposes varying utilization requirements as shown in appendix B.

III. Mathematical Model

In this section we define a space-time network and formulate a linear program on this network that trades off load-balancing and communication costs of energy consumption and delay.

A. Space-Time Network

To capture the dynamics of the network over consecutive time epochs, we take snapshots of the changing network topology at $\tau = \frac{F}{B}$ consecutive units of time over the horizon $\mathcal{T}$. In this way a time expanded network is formed where the position of all mobile nodes is described by the coordinate (i, a) indicating the position of node $i \in \mathcal{N}$ at the a^{th} time epoch. Clearly, for all static nodes, the time coordinate is redundant and thus their positions can be uniquely identified by the node index i. Illustratively, figure 3 shows the time expanded network for a single user, two BSs and four mobile relay nodes at three consecutive time intervals. At time epoch 0 the current position of all nodes in the network is shown. Also a mobile node $i \in \mathcal{N}$ at the first time epoch is index as $(i, 1)$ while for the second time epoch is $(i, 2)$ and so on. Set V_p contains all space-time network nodes of the form (i, a), where $i \in \mathcal{N}$ at the a^{th} time epoch. Future node positions can be generated by mobility prediction models to assess the network conditions at different scenarios. It is important to note at this stage that we do not require accurate mobility predictions for the proposed scheme to operate efficiently. In fact, we only

need to know the candidate serving BSs of a node and a rough estimate of its location for each time epoch.

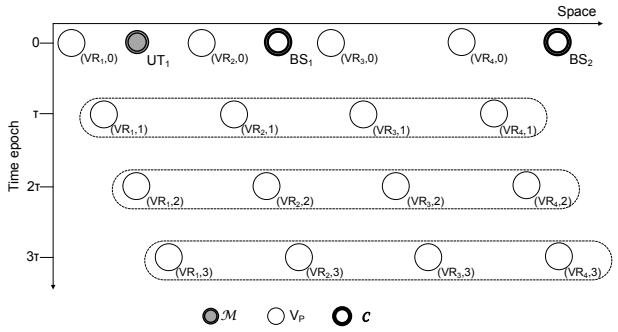

Fig. 3: Illustration of an instance of the space-time network under consideration. Future mobile node positions are replicated at consecutive time epochs to generate a static network over time.

B. Graph on the Space-Time Network

On the space-time network described in the previous subsection, we define the graph $G = (V, L)$ where $V = \{\mathcal{M}, V_p, \mathcal{C}, BSC\}$ is the set of nodes and L is the set of links, where we introduce node BSC (denoted by K) as a super sink node for all user requests. Node BSC can be considered as a physical entity (as is the case with current cellular deployments that have a radio network controller) or a virtual node acting as a load aggregator. Further, the links in L are subdivided into three distinct subsets described as follows:

- L_1 : Set of links where transmission between a transceiver pair takes place.
- L_2 : Set of links where no transmission occurs and information is physically propagated by mobile nodes.
- L_3 : No information travels on these links. They are dummy links that connect all BSs to the super sink node BSC to ensure load balancing.

The L_1 links are of the form $i \mapsto j$, $i \in \mathcal{M} \cup V_p, j \in V_p \cup \mathcal{C}$, where transmission of a single message occurs between nodes i and j. For example, links $UT_1 \mapsto BS_1$, $UT_1 \mapsto (VR_2, 2)$, $(VR_2, 1) \mapsto (VR_3, 2)$ and $(VR_3, 3) \mapsto BS_2$ shown in figure 4 are all L_1 links. Notice that node replication on the space-time network is done every τ units of time. Therefore, for all communication links (i.e. links in L_1) a single message can be transmitted at consecutive time instances. The delay in communication for links in L_1 can vary with respect to the type of L_1 link. There are four types of L_1 links shown below with their delays:

$$\begin{array}{lll} \text{Case } 1: & UT_m \mapsto BS_k & \text{delay } \tau \\ 2: & UT_m \mapsto (VR_n, a) & \text{delay } a\tau \\ 3: & (VR_n, a) \mapsto (VR_l, a+1) & \text{delay } \tau \\ 4: & (VR_n, a) \mapsto BS_k & \text{delay } \tau \end{array} \quad (17)$$

where $m \in \mathcal{M}$, $k \in \mathcal{C}$, $n, l \in \mathcal{N}$ and $a \in \mathcal{T}$. The first case is simply the traditional direct communication and the delay is merely the communication time τ. For example, the first case is link $UT_1 \mapsto BS_1$ as shown in figure 4. In the second case, the message leaves the user at time 0 and arrives at time epoch a; initially the user for the first $a-1$ time epochs is buffering the message while waiting for a mobile node to approach and in the last time epoch is transmits the message. So the total delay is $a\tau$. An example of this in figure 4 is link $UT_1 \mapsto (VR_2, 2)$ where the message is buffered for the first time epoch and transmitted in the second time epoch. The third case represents transmission between two mobile nodes and incurs delay τ. In figure 4 this case is represented by link $(VR_2, 1) \mapsto (VR_3, 2)$. The fourth case represents transmission from a mobile node to a base station as shown by link $(VR_3, 2) \mapsto BS_2$ in figure 4.

Links in L_2 is denoted as follows, $L_2 = \{(i, a) \mapsto (i, a+1) : \forall (i, a), (i, a+1) \in V_p\}$ where a mobile node carries messages from one time period to the next without any transmission taking place. For example in figure 4, $(VR_2, 1) \mapsto (VR_2, 2)$ and $(VR_3, 2) \mapsto (VR_3, 3)$ are two examples of L_2.

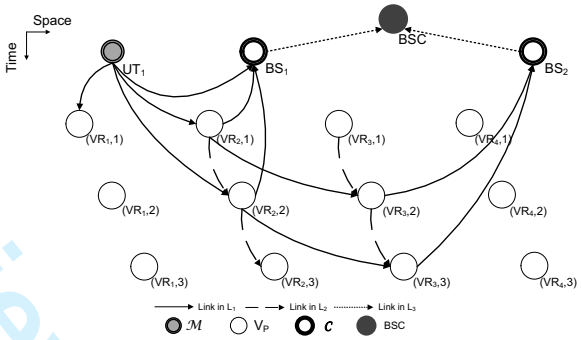

Fig. 4: The figure illustrates a connected graph for the space-time network under consideration. Selectively a number of links are shown.

Links in L_3 are dummy links that we place between the base stations and the sink node K in order to achieve load balancing. Since a BS can accept at most u_k requests, similar to the network of problem (P2) (figure 2), we place multiple arcs between BS k and sink node K each indexed by $n = 1, \ldots, u_k$ of increasing convex cost to promote load balancing as in the problem P_2. Therefore, link $(kKn) \in L_3$, $k \in \mathcal{C}, n = 1, \ldots, u_k$ identifies the n^{th} link emanating from BS $k \in \mathcal{C}$ and ending at the BSC. The cost on these links expresses the increase in utilization of accepting user requests through an arbitrary BS. Therefore, the function $w_k(n)$ as defined by equation (10) determines the cost of accepting the n^{th} message through BS k.

We define the flow variable x_{ij} for links in $L_1 \cup L_2$ and flow variable y^n_{kK} for links in L_3. Further, for links in L_3 the flow should be kept binary to indicate the accept/reject decision of each request. It also accounts for the increase in load by an arbitrary BS when accepting a user request as described in section II-B. On the other hand, mobile nodes can buffer an arbitrary number of messages while in transit. Therefore, the capacity for all links in L_1 is $u_{ij} = 1$ and for links in L_2 we set it to the maximum possible, which is $u_{ij} = M$, since there are M users with one message each. In addition, the capacity for links in L_3 is 1 unit flow as detailed in section II-B. We define the supply and demand

for all nodes $i \in V$ as follows:

$$b(i) = \begin{cases} +1 & \text{i} \in \mathcal{M} \\ -M & \text{i} = \text{K} \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

C. Link costs

The energy consumed in transmission of a unit flow for links in L_1 and L_2, is given as follows:

$$E(i \mapsto j) = \begin{cases} f_{ij} & for\ i \mapsto j \in L_1 \\ 0 & for\ i \mapsto j \in L_2, \end{cases} \quad (19)$$

where f_{ij} is defined in section II-A. Note that in equation (2) the energy consumed in transmission is correct only for a unit flow as it is not linear to the number of flows send. However, for the multiplicative cost $x_{ij}f_{ij}$ on a link in L_1 as expressed by equation (19) it gives the correct cost as it can attain only binary values.

The delay per unit flow for all links in $L_1 \cup L_2$, using $j = (k, a+1)$, is as follows:

$$\Phi(i \mapsto j) = \begin{cases} \tau & for\ i \mapsto j \in L_1 \cup L_2,\ i \notin \mathcal{M} \\ \tau(a+1) & for\ i \mapsto j \in L_1,\ i \in \mathcal{M}, \end{cases} \quad (20)$$

where for link $i \mapsto j \in L_1$, $i \in \mathcal{M}$, the user postpones transmission for a time periods before transmitting. The total communication cost is thus defined as a weighted sum of the communication energy consumption and the delay incurred while traversing the link and is defined as follows, $c_{ij} = E(i \mapsto j) + \lambda\Phi(i \mapsto j)$, $i \mapsto j \in L_1 \cup L_2$ where λ is the weighting parameter for the two quantities. Note that for hard deadline constraints, node replication can be restricted to a finite horizon $\mathcal{T}$ that allows for the maximum delay tolerance of the service considered. In the latter case, the model is still valid, however the cost of L_1 links is simply the energy consumption as expressed in equation (19), while for the L_2 links the traversal cost is zero.

For links in L_3, both the energy consumption and communication delay are assumed to be negligible.

D. Mathematical Programming Formulation

The space-time network model derived in the previous section generates all feasible paths from the source nodes to the BSC through the available BSs. Clearly, decisions need to be made on the best forwarding nodes that could potentially forward messages and the time at which these nodes need to forward data. In this section, we formulate the joint routing and scheduling problem to generate the optimal decision policies for message forwarding. We are interested in finding forwarding paths that minimize the load imbalance in the system and at the same time reducing the communication cost of the forwarding paths. The following linear integer mathematical program is formulated where the weights between the competitive parameters of load balancing, communication energy consumption and message delivery delay can be set using parameters λ and γ. These parameters can be weighted accordingly depending on network operation preferences.

$$\text{(P3)}\ \text{minimize} \sum_{i \mapsto j \in L_1 \cup L_2} c_{ij}x_{ij} + \gamma \sum_{(kKn) \in L_3} w_k(n) y^n_{kK} \quad (21)$$

$$\text{s.t.} \sum_{j: i \mapsto j \in L_1} x_{ij} \leq 1\ \forall\, i,\ i \in V_p \quad (22)$$

$$\sum_{j: j \mapsto i \in L_1} x_{ji} \leq 1\ \forall\, i,\ i \in V_p \quad (23)$$

$$\sum_{j: i \mapsto j \in L_1 \cup L_2} x_{ij} - \sum_{j: j \mapsto i \in L_1 \cup L_2} x_{ji} = b(i)\ \forall\, i,\ i \in \mathcal{M} \cup V_p \quad (24)$$

$$\sum_{n:(kKn) \in L_3} y^n_{kK} - \sum_{j: j \mapsto k \in L_1} x_{jk} = b(k)\ \forall\, k,\ k \in \mathcal{C} \quad (25)$$

$$- \sum_{n:(kKn) \in L_3} y^n_{kK} = b(K) \quad (26)$$

$$0 \leq x_{ij} \leq u_{ij},\ x_{ij} \in \mathbb{Z},\ y^n_{kK} \in \{0, 1\} \quad (27)$$

The objective function of problem (P3) minimizes the weighted sum cost of load imbalance and communication cost in the multi-cell network topology. Constraint equations (22-23) restrict all relay nodes to transmit and receive from a single other node at any one time. Constraints (24-26) are the flow conservation equations guaranteeing the generation of the end-to-end paths. Equations (27) are the capacity and integrality constraints on the flow variables. Integrality is a natural assumption as all information messages are considered to be independent units that can not be split.

Due to the integer variables the mathematical program formulated by equations (21-27) is in general hard to solve. However, it can be shown that the linear programming relaxation of problem (P3) guarantees to give integer solutions. Therefore, the problem can be solved efficiently in time even for large instances.

Theorem 1: Let $A[x; y] \leq z$ be the matrix representation of constraints (22)-(27), ignoring integrality constraints. Then A is totally unimodular (TU).

The proof is detailed in appendix C. It is well known that if A is TU and z is integer then the feasible region described by $A[x; y] \leq z$ has integer extreme points. Thus, we can relax the integrality constraints and still get integer solutions using linear programming (simplex method). This means that we can solve the above mathematical program for a large number of nodes in very short running times.

IV. Numerical Investigations

Solving problem (P3) to optimality for different values of the weighting parameters, we study in this section the possible performance gains of the proposed generalized SCF scheme.

A. Simulation set-up

For the communication model we assume that all source nodes have a single message of size $F = 4$Mbit to

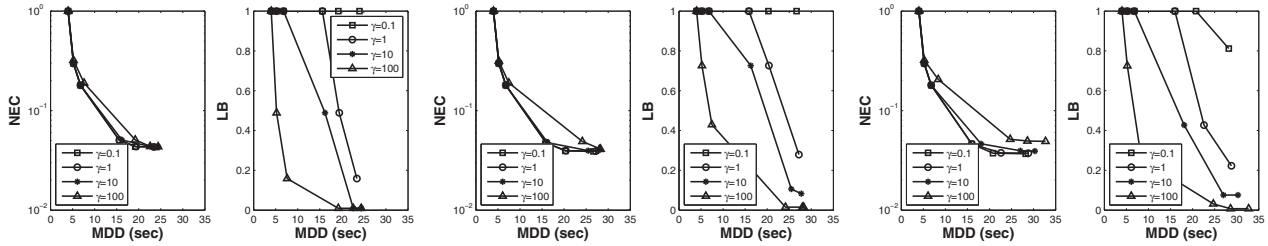

Fig. 5: Pareto curves for the optimal operating points of the proposed SCF relaying scheme, where γ is the tradeoff coefficient of load balancing versus energy efficiency and delay: for a network topology of (a) 2 cells, 20 users and 40 relay nodes, (b) 3 cells, 30 users and 60 relay nodes and (c) 4 cells, 40 users and 80 relay nodes.

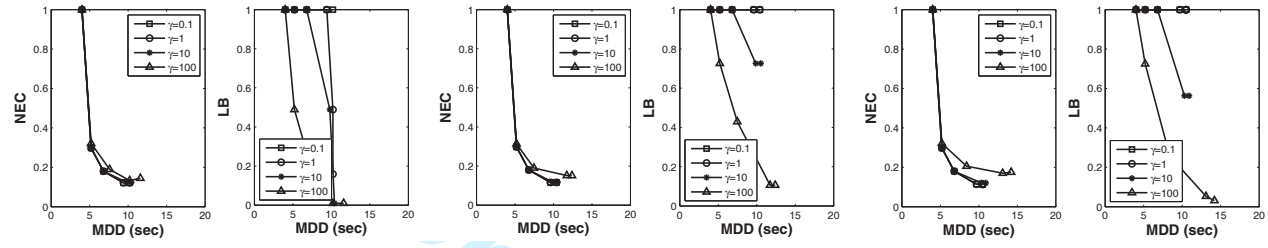

Fig. 6: Performance evaluation of the basic multihop scheme: for a network topology of (a) 2 cells, 20 users and 40 relay nodes, (b) 3 cells, 30 users and 60 relay nodes and (c) 4 cells, 40 users and 80 relay nodes.

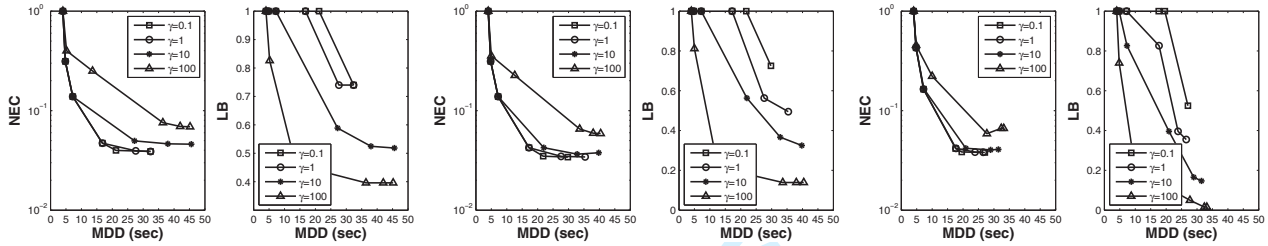

Fig. 7: Effect of the data request volume on the system performance gains: for a network topology of (a) 4 cells, 10 users and 80 relay nodes, (b) 4 cells, 20 users and 80 relay nodes and (c) 4 cells, 30 users and 80 relay nodes.

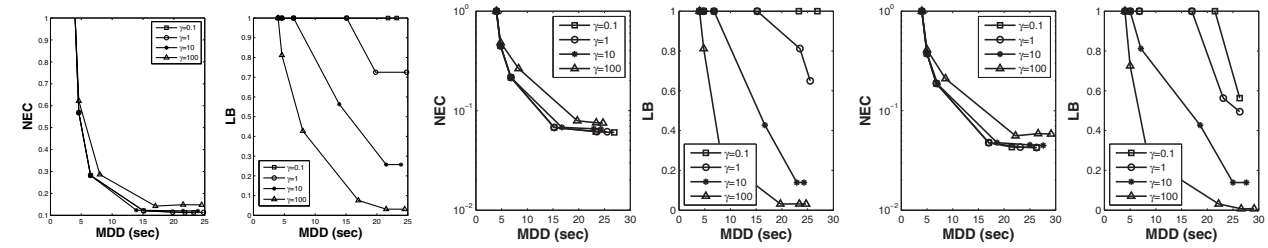

Fig. 8: The effect of mobile node density on system performance: for (a) 4 cells, 40 users and 10 relay nodes, (b) 4 cells, 40 users and 20 relay nodes and (c) 4 cells, 40 users and 30 relay nodes. Clearly for few relay nodes, the energy efficiency gains reduce as the buffering queues grow larger. However, even for low relay node densities, load balancing gains are intact.

communicate to the wired network. All links transmit at a rate of $B = 1$Mbps and the radio wavelength is assumed to be $\nu = 0.126$m. The antenna height for all nodes is assume to be $h_v = 1.5$m, except for the BS where it is $h_b = 15m$. Even though the device-to-device and device-to-BS communications possess very different propagation characteristics, we use the same channel model. A more detailed channel model is not necessary since this work is concerned with the energy efficiency gains that are achieved from communicating shorter distances.

The power consumed by the circuit electronics during the transmit/receive operations is $P_c = 50$mW, while the received power threshold for successful communication is $P_r = -52$dBm[1]. From these values, the following parameters can be deduced: The threshold distance for switching between the two modes of propagation can be estimated as follow, $d_{brake} = \frac{4\pi h^2}{\nu}$. The energy per bit consumed by the power amplifier is $e_d(los) = \frac{P_r(4\pi)^2}{B\nu^2}$

[1]This is similar to the average signal strength for an LTE receiver at 10MHz channel bandwidth which is calculated in [20] to be -56.5dBm.

for free space losses with pathloss exponent $\eta = 2$ and $e_d(mp) = \frac{P_r}{Bh^4}$ for the plane earth model with pathloss exponent $\eta = 4$. The fixed circuit energy consumption is $e_t = e_r = \frac{P_c}{B}$.

A network topology of $C \in \{2, 3, 4\}$ cells is considered with a cell radius of $R = 400$m. The initial cell loading levels follow a truncated normal distribution in the interval 0 - 60 Mbps with variance $\sigma^2 = 0.5$. Without loss of generality, we assume that all user requests in the network can be served by each BS. Alternatively, the capacity of each BS can be set appropriately as described in section III. The resource consumption level for all users is simply approximated to the data rate, q(F)=B. $M \in \{10, 20, 30, 40\}$ uniformly distributed source nodes reside in each cell and $N \in \{20, 30, 40\}$ mobile nodes travel along each direction of a stretch of road covering the network diameter. All transmissions are assumed to be restricted to a maximum range of R meters.

The popular open-source microscopic mobility simulator (SUMO - www.dlr.de/ts/sumo) has been employed to test realistic mobility scenarios. In addition, Matlab (www.mathworks.com) was used to set up the simulation, in which the space-time network was build and the networking conditions were developed for each scenario studied (including the definition of all variables, the computation of all necessary values and the logging of the results from the optimization problems). The Gurobi optimization solver (www.gurobi.com) was used to compute the optimal relay strategies based on the proposed mathematical programming formulations. The car-following parameters used within SUMO for all vehicular nodes are as follows: maximum acceleration is 0.8m/s^2, maximum deceleration is 4.5m/s^2, maximum travelling speed 14m/s, the car length is 5 meters and response time to unpredictable events is set to 0.5 seconds.

Simulations were repeated 1000 times for each setting to obtain statistically unbiased results.

B. Simulation results in general topologies

Figure 5 plots the optimal energy-delay tradeoffs for different values of the weighting parameters. The figure on the left depicts the normalized communication energy consumption (NEC) against the message delivery delay (MDD) while the figure to the right shows the normalized load balancing (LB) improvements versus message delivery. The LB metric indicates the variance in data traffic served across the cells (after decisions are made on which requests are going to be served by each cell). As expected, for increased delivery delays, more forwarding paths become feasible and thus messages can be forwarded to neighboring cells to reduce the load imbalance (i.e., the data traffic variance between the cells). More importantly however is the fact that for increased message delivery delays, the forwarding paths can achieve the minimum energy values. As shown in figure 5, when enough delay can be tolerated on the end-to-end paths, the minimum energy cost paths can be attained for various target load balancing conditions. On the other hand, when higher weight is given to load balancing instead of energy efficiency (i.e., higher values of the γ parameter) then for a target MDD the minimum load imbalance is achieved at the expense of higher energy consumption. In effect, all those paths that are able to steer traffic away from hotspots and into neighboring cells are chosen to minimize the load imbalance but due to the time restrictions in the delivery delay transmissions may occur at longer distances causing a rise in energy consumption. Hence, a delay flexibility on the forwarding paths allows for both the energy efficiency gains and load balancing improvements. Note that the benefits from tolerating delays of few seconds are significant; in fact Fig. 5a shows that for a small delay of 25sec, the load imbalance between all cell (for all scenarios considered) drops my an order of magnitude while the normalised energy consumption drops from 1 to 0.03 offering in that respect a gain of approximately $33\times$. Also it is important to note that the benefits of considering increasingly many cooperating BSs are only incremental for both the energy efficiency gains and the load balancing improvements (cooperation of 2, 3 and 4 cells are shown in figures 5a, 5b and 5c respectively).

Further, the performance of the proposed generalized multihop scheme that incorporates SCF relaying is compared to the basic multihop (BMH) scheme where messages are forwarded by the nodes as soon as they are received and no delay on forwarding is allowed. In figure 6 the energy delay trade-offs for the basic multihop scheme are shown. Clearly, BMH strives to achieve the same load balancing improvements compared to SCF however at an increase communication cost. Notice that the energy consumption in this case drops by merely an order of magnitude. Comparing Figs. 5a and 6a, for a delivery delay of 15 sec the load balancing performance between the two schemes is the same but in the case of the proposed SCF scheme, the drop in energy consumption is significantly greater with a drop to 0.03 observed as opposed to a drop by 0.8 observed by BMH. Evidently this is due to the fact that the basic multihop scheme does not take advantage of the increased delay tolerance of elastic services and thus the delay is bounded to the retransmission delay of the en route hops. Inevitably in this case, the communication distance is higher and thus load balancing improvements are achieved at the expense of communication cost, in contrast to the proposed scheme.

Clearly, the availability of delay tolerant source nodes allows for higher improvements on load balancing. Figure 7 illustrates the effect of increasing higher number of subscriber base with delay tolerant traffic in the system. Looking into the performance improvements of considering 10, 20 and 30 active source nodes in a network topology of $C = 4$ cells in figure 7, considerable improvements on load balancing can be obtained for delay tolerant message deliveries. Note also that the communication cost is not affected by the increased number of source nodes considered. On the other hand, the effect on the communication cost is more severe when the number of source nodes is sufficiently larger than the candidate mobile forwarding nodes. As

seen in figure 8, for a fixed number of source nodes, the availability of mobile relay nodes considerably affects the communication cost. This is expected as the mobile nodes receive messages from more source nodes which in turn transmit over longer distances when their queues have higher loads. As the number of candidate forwarding nodes increases however, there are more forwarding possibilities and the load is more evenly distributed across the nodes. Importantly though, is the fact that even for instances when few mobile nodes exist the load balancing improvements are not deteriorated. This is so since mobile nodes act as message ferries; collecting and delivering delay tolerant messages across adjacent cells.

Unfortunately, there are no general rules for choosing the data routes since they change according to the many factors most important of which are the data traffic load of each cell, the delay tolerance of the application, the opportunities for SCF relaying provided by vehicle traffic routes and available vehicles. That is why we use the optimal solution of an optimisation problem, which can tackle realistic size networks in tractable computational times.

C. Simulation results in realistic topologies

A realistic network topology is considered in this section; a segment of a cellular network near King's College London (London, UK) is investigated with two BSs of a single cellular operator[2]. Further, the topology layout is imported from OpenStreetMap[3] into SUMO where mobility traces are generated for the marked roadway shown in figure 9. Note that along this route, 4 traffic light junctions are operating. For the simulation, $M = 20$ active users are assumed to be active along this route while $N = 20$ vehicles travel in either direction of the roadway. Here we assume that the micro-cells have cell radius of $R = 200$m while all other parameters are as described in section IV. As before, Matlab was coupled with SUMO to run the simulations and the Gurobi optimization solver was used to compute the optimal forwarding decisions based on the proposed formulation.

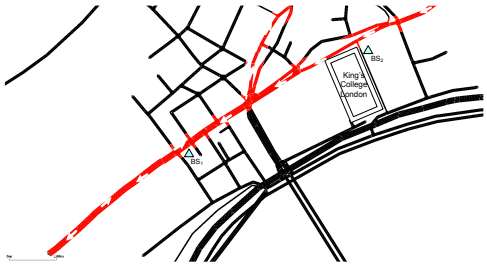

Fig. 9: Network deployment in the area outside the Strand Campus at King's College London. Two BSs from a single UK operator are shown in the figure.

In figure 10 the optimal energy-delay and load-delay curves are presented. As expected, the performance improvements are considerable when increased delay on message delivery is tolerable. There are a few things worth to be noted here. First of all to achieve these gains note that there is an increase end-to-end delay. This is due to the fact that there are frequent stops of vehicles controlled by the roadside traffic lights. This results in vehicle buffering messages over longer periods. Secondly, to achieve the minimum load imbalance in the system at low delivery delays there is increase communication cost as more retransmissions are needed to shift data traffic between the cells. However, when elastic services are considered, then the communication cost of the SCF scheme is always less than the direct link cost. Thirdly, all transmission links experience free space losses with a path-loss exponent of $n = 2$ (this result is based on the model parameters derived in section IV where transmissions in ranges less than 200m are considered line-of-sight links). Therefore, the energy efficiency gains presented here can be considered as a lower bound on the actual gains expected in a real world implementation.

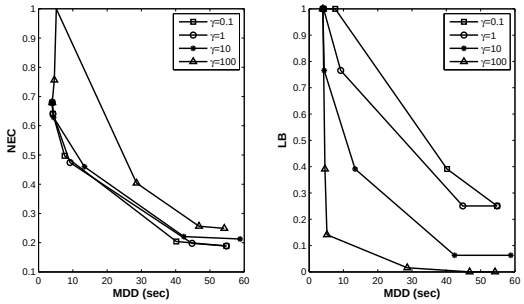

Fig. 10: Energy-delay and load balancing-delay curves for a real network topology.

Nevertheless, the proposed network model can be used in practice to derive decision policies based on the desired look-ahead strategies to be implemented. As shown in section III the mathematical programming formulation for the optimal routing and scheduling decisions is computationally efficient and thus can be used in practical implementations. Further, when the prediction accuracy of mobile terminal positions can not be assumed for a large time horizon, then the problem can be solved iteratively at consecutive time steps to correct the forwarding decisions been made.

V. Practical Considerations

In a practical setting an iterative scheme can be employed in which the forwarding decisions are (re)computed over a moving time horizon. At first, decisions are made over time horizon T as defined in section III. At subsequent stages, updated vehicle locations and supply/demand parameters are used to reconstruct the space-time network which is used to re-compute the forwarding decisions. This iterative procedure can continue indefinitely.

Algorithm 1 describes this iterative procedure when recomputations are made at a rate θ. Initially, the most recent node location information is used (step 1) and the upload requests (step 2), are used to construct the space-time

[2]Base station positions are extracted from http://www.sitefinder.ofcom.org.uk/.

[3]http://www.openstreetmap.org/

Algorithm 1 Iterative-SCF scheme.

Ensure: k=0.

1: Update vehicle positions based on newly received vehicle location information.
2: Update supply/demand parameters.
3: Re-construct space time network using steps 1, 2.
4: Solve problem (P1) for time horizon $t = k\tau, k\tau + 1, \ldots, k\tau + T$.
5: Execute decisions for the first θ time periods.
6: $k = k + \theta$; Go to step 1.

network (step 3). Then, the forwarding decisions are derived (step 4) for the next T time periods, but only the decisions for the next θ time periods are executed (step 5) before re-computing the forwarding policies. Notably, the proposed message forwarding scheme relies on knowledge of the underlying network dynamics to compute the forwarding decision. Of course, accurate terminal location information has become an increasingly valuable commodity (not only for location based services [34] [35] [36] but also for network optimization [37]) and has been an issue of significant standardization efforts recently, [38] [39]. Not only that, network operators have been increasingly interested in leveraging the terminal capabilities to collect statistics of the underlying network conditions. The document in [40] details the logging and reporting procedures implemented at the user terminals to obtain such information from the surrounding environment. Table II below, details the measurement fields standardize by those mechanisms. It is important to note here that such information presents a complete set of measurements required for the proposed solution as well.

TABLE II: Location information overhead

Parameter	Size (bits)	Definition
Location	63	Lat/ Lon/ Alt information
Time stamp	40	Month/ Day/ Hour/ Min/ Sec
CGI	52	Serving cell id
PCI	288	Neighboring cell id (x32)
Measurements	429	Radio environment measurements

VI. Conclusions

The key contribution of the work is the explicit amalgamation of delay tolerant techniques with cellular networks especially as pertain to the task of load balancing. However, despite the reported gains the main limitation is that a new protocol stack on top of TCP/IP is required in order to handle large delays. Most previous studies on load balancing schemes assume real-time service constraints but these restrictions can be relaxed when elastic Internet traffic is considered. Hence, for elastic services, the load imbalance and energy efficiency can significantly be improved by allowing mobile nodes to shift traffic from hot spots to neighbouring less congested cells in a store-carry and forward manner. Hence, not only edge cell users can contribute to the load balancing improvements but also mobile nodes at arbitrary cell locations.

A mathematical programming model is derived that is computationally efficient and can be used in practical settings to find optimal routing and scheduling policies. Alternatively, this can be considered as a bound on the performance that can be achieved by other greedy/heuristic mechanisms that use local information. Considering both the communication cost of the en-route paths and the load balancing improvements in the network, we show that the SCF scheme is persistently preferred over the alternative routes for delay insensitive traffic. For the latter case, we demonstrate that considerable load balancing improvements and reduced energy cost can be achieved via SCF relaying compared to the single hop and standard multihop routes.

Acknowledgment

The work reported in this paper has formed part of the Green Radio Core 5 Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com. This research has been funded by EPSRC and by the Industrial Companies who are Members of Mobile VCE.

References

[1] M. Jung-min and C. Dong-ho, Efficient Cell Selection Algorithm in Hierarchical Cellular Networks: Multi-User Coordination, *IEEE Communications Letters*, Vol. 14 , No. 2, Feb. 2010, Page(s):157 - 159.
[2] S. Kyuho, C. Song and G. de Veciana, Dynamic Association for Load Balancing and Interference Avoidance in Multi-cell Networks, *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Apr. 2007.
[3] H. Kim, G. de Veciana, Y. Xiangying and M. Venkatachalam, α-Optimal User Association and Cell Load Balancing in Wireless Networks, *IEEE International Conference on Computer Communications*, Mar. 2010.
[4] Y. Bejerano, H. Seung-Jae and Li Li, Fairness and Load Balancing in Wireless LANs Using Association Control, *IEEE/ACM Transactions on Networking*, June 2007, Page(s):560 - 573.
[5] Y. Bejerano, H. Seung-Jae, Cell Breathing Techniques for Load Balancing in Wireless LANs, *IEEE Transactions on Mobile Computing*, Vol. 8 , No. 6, June 2009, Page(s):735 - 749.
[6] K.A. Ali, H.S. Hassanein and H.T. Mouftah, A Novel Dynamic Directional Cell Breathing Mechanism with Rate Adaptation for Congestion Control in WCDMA Networks, *IEEE Wireless Communications and Networking Conference*, Mar. 2008.
[7] P. Bahl, et. al, Cell Breathing in Wireless LANs: Algorithms and Evaluation, *IEEE Transactions on Mobile Computing*, Vol. 6, No. 2, Feb. 2007, Page(s):164 - 178.
[8] O. Tonguz and E. Yanmaz, The Mathematical Theory of Dynamic Load Balancing in Cellular Networks, *IEEE Transactions on Mobile Computing*, Vol. 7, No. 12, Dec. 2008, Page(s):1504 - 1518 .
[9] A. Sang, et. al, Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems, *Wireless Networks*, Vol. 14, No. 1, Feb. 2008, Page(s): 103 - 120.
[10] K. Papadaki and V. Friderikos, Optimal Vertical Handover Control Policies for Co-operativeWireless Networks, *Journal of Communications and Networks*, Vol. 5, No. 3, Dec. 2006, Page(s):442 - 450.
[11] K. Fall and S. Farrell, DTN: An Architectural Retrospective, *IEEE Journal on Selected Areas in Communications*, Vol. 26, No. 5, Jun. 2008, Page(s):828 - 836.
[12] C. Chen and Z. Chen, Exploiting Contact Spatial Dependency for Opportunistic Message Forwarding, *IEEE Transactions on Mobile Computing*, Vol. 8, No. 10, Oct. 2009, Page(s):1397 - 1411.
[13] M. Grossglauser and D. Tse, Mobility Increases the Capacity of Ad Hoc Wireless Networks, *IEEE/ACM Transactions on Networking*, Vol. 10, No. 4, Aug. 2002, Page(s):477 - 486.
[14] S.C. Borst, N. Hegde and A. Proutiere, Mobility-Driven Scheduling in Wireless Networks, *IEEE Conference on Computer Communications*, Apr. 2009, Page(s):1260 - 1268.
[15] S. Chakraborty, Y. Dong, D. Yau and J. Lui, On the Effectiveness of Movement Prediction to Reduce Energy Consumption in Wireless Communication, *IEEE Transactions on Mobile Computing*, Vol. 5, No. 2, Feb. 2006, Page(s):157 - 169.
[16] Y. Dong, WK. Hon, D. Yau and JC. Chin, Distance Reduction in Mobile Wireless Communication: Lower Bound Analysis and Practical Attainment, *IEEE Transactions on Mobile Computing*, Vol. 8, No. 2, Feb. 2009, Page(s):276 - 287.

[17] A. Venkateswaran, V. Sarangan, T. La Porta and R. Acharya, A Mobility-Prediction-Based Relay Deployment Framework for Conserving Power in MANETs , *IEEE Transactions on Mobile Computing*, Vol. 8, No. 6, Jun. 2009, Page(s):750 - 765.
[18] C. Grote, IoT on the Move: The Ultimate Driving Machine as the Ultimate Mobile Thing, *IEEE International Conference on Pervasive Computing and Communications*, 2014.
[19] R.K.Ahuja, T.L.Magnanti and J.B.Orlin, Network Flows: Theory, algorithms, and applications, *Prentice Hall*, 1993, Page(s):318.
[20] H. Holma and A. Toskala, LTE for UMTS - OFDMA and SC-FDMA Based Radio Access, *John Wiley & Sons, Ltd*, 1st. Ed., 2009, Page(s):341.
[21] P. Kolios, V. Friderikos and K. Papadaki, Energy-aware mobile video transmission utilizing mobility, *IEEE Network*, Vol. 27, No. 2, Mar. 2013.
[22] G. Alyfantis, S. Hadjiefthymiades and L. Merakos, Exploiting user location for load balancing WLANs and improving wireless QoS, *ACM Transactions on Autonomous and Adaptive Systems*, Vol. 4, No. 2, May 2009.
[23] A. Balachandran, P. Bahl and G. M. Voelker, Hot-Spot Congestion Relief in Public-Area Wireless Networks, *IEEE Workshop on Mobile Computing Systems and Applications*, Aug. 2002.
[24] G. Fettweis and E. Zimmermann, ICT Energy Consumption - Trends and Challenges, *International Symposium on Wireless Personal Multimedia Communications*, Sept. 2008.
[25] S. Vadgama, Trends in Green Wireless Access, *Fujitsu Sci. Tech. J.*, Vol: 45, No: 4, Oct. 2009, Page(s):404 - 408.
[26] Cisco, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012 - 2017, *Cisco white paper*, Feb. 2013.
[27] P. Kolios, V. Friderikos and K.Papadaki, Energy Efficient Relaying via Store-Carry and Forward within the Cell, *IEEE Transactions on Mobile Computing*, Nov. 2012.
[28] NGMN Alliance, Next Generation Mobile Networks 5G White Paper V1.0, *NGMN 5G Initiative, February 2015*
[29] 3GPP RWS-120010 WS Docomo, Requirement, candidate Solutions and Technology Roadmap for LTE Rel-12 onward, *3GPP Workshop on Release 12 and on-wards Ljubljana, Slovenia, June 11-12, 2012*
[30] 5G Radio Access: Requirements, Concepts and Technologies, *Docomo 5G White Paper, July 2014*
[31] L. Chen, L. Libman, and J. Leneutre, Conflicts and incentives in wireless cooperative relaying: A distributed market pricing framework. IEEE Transactions on Parallel and Distributed Systems, Vol. 22, No. 5, May 2011, Page(s): 758-772.
[32] C. Comaniciu, B.M. Narayan, H. Vincent Poor, and J.M. Gorce, An Auctioning Mechanism for Green Radio, Journal of Communications and Networks, Vol. 12, No. 2, April 2010, Page(s):114 - 121.
[33] A. Schrijver, Theory of Linear and Integer Programming, *Wiley Press*, April 1998.
[34] S. Wang, J Min and B. Yi, Location Based Services for Mobiles: Technologies and Standards, *IEEE International Conference on Communications*, Tutorial, May 2008.
[35] Google, The Mobile Movement: Understanding Smartphone Users, *Google/IPSOS OTX MediaCT*, Apr. 2011, www.google.com/think/insights.
[36] C. Botezatu and C. Barca, Intelligent vehicle safety-eCALL, *e-Society*, 2008.
[37] T. Jansen, et al. Handover parameter optimization in LTE self-organizing networks, *IEEE Vehicular Technology Conference*, Sept. 2009.
[38] Y. Zhao, Standardization of Mobile Phone Positioning for 3G Systems, *IEEE Communications Magazine*, Vol. 40, No. 7, Jul 2002, Page(s):108 - 116.
[39] ETSI 3rd Generation Partnership Project, LTE; Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Stage 2 functional specification of user equipment positioning in E-UTRAN, *ETSI Technical Specification 136 305 V9.0.0 Release 9*, Oct. 2009.
[40] 3GPP, Technical Specification Group Radio Access Network Study on Minimization of drive-tests in Next Generation Networks, *3GPP*, TR 36.805 v9.0.0, Dec. 2009.

APPENDIX

Appendix A

Load variance minimization: proof of proposition 2: Proof:

Concerning the objective function of problem (P1): Expanding equation (3) and using $\sum_{k=1}^{C} y_k = M$, which ensures that all user requests are accepted by the network, results in the following expression:

$$\text{Var} = \sum_{k=1}^{C} (y_k q)^2 + 2 \sum_{k=1}^{C} (y_k q)\, S_k + \mathfrak{f}(M, S_k, q) \quad (28)$$

where the function $\mathfrak{f}(M, S_k, q)$ does not depend on y_k.

Concerning the objective of problem (P2): Let $v_k = \sum_{n=1}^{u_k} y_{kK}^n$ be the number of requests satisfied by BS k. The weights of decision variables y_{kK}^n in the objective for $n = 1, \ldots, u_k$ satisfy $w_k(1) < \ldots < w_k(u_k)$ and since we are minimizing, the optimal solution will choose the first v_k arcs. Thus we can write the objective of (P2) as follows:

$$\sum_{k=1}^{C} \left[\sum_{n=1}^{v_k} (w_k(n)) \right] = \sum_{k=1}^{C} \left[v_k q S_k + \frac{v_k [v_k + 1] q^2}{2} \right] \quad (29)$$

As a sum of its terms, equation (29) can be expressed as follows:

$$\sum_{k=1}^{C} v_k q S_k + \frac{1}{2} \sum_{k=1}^{C} [v_k q]^2 + \frac{q^2}{2} \sum_{k=1}^{C} v_k \quad (30)$$

Since $\sum_{m=1}^{C} v_k = M$, then the last term in (30) is a constant.

From (28), we can see that (30) is equivalent to $\frac{\text{Var}}{2}$. Thus, the objective functions of problems (P1) and (P2) are equivalent.

Consider now the constraints in both (P1) and (P2). It is straightforward to see that $v_k = \sum_{m=1}^{u_k} y_{kK}^m$ in (P2) is identical to y_k in (P1). In effect, constraint (5) is identical to (11), (6) is identical to (12), (7) is identical to (15) and (8) is identical to (14). Moreover, constraint (13) in (P2) is redundant and thus the two problems are equivalent. ■

Appendix B

Load balancing model extensions: The derivation of the load balancing model in section II-B considered the case when all user requests consumed the same amount of resources by all infrastructure nodes. Here, we extend that model to consider the case when user requests require different resource consumption levels. For the model structure and the formulation hereafter consider the illustration in figure 2. Let q_m be the resource consumption level of the m^{th} request. Then, the greatest common divisor of all request levels q_m, $m \in \mathcal{M}$ is $\mathfrak{q}$. Given $\mathfrak{q}$, the capacity of all links $m \mapsto k, \forall m \in \mathcal{M}, \forall k \in \mathcal{C}$ can be expressed as $u_k^m = \frac{q_m}{\mathfrak{q}}$. Further, here we have two types of variables, the flow variables z, y as shown in the diagram in figure 2 and binary variables δ indicating the acceptance (or not) of a request through a specific infrastructure node.

As before the flow variable from a source node m to BS k is z_k^m and the cost of the flow is assumed to be zero. Further, the remaining capacity of the BSs is decomposed into the $\mathfrak{q}$ parts as defined above and thus the total number of links emanating from k and ending at K is $\Delta_k = \min(\sum_{m=1}^{M} \frac{q_m}{\mathfrak{q}}, \left\lfloor \frac{U_k - S_k}{\mathfrak{q}} \right\rfloor)$ (according to figure 2, $u_k = \Delta_k$). A flow variable flow y_{kK}^i, $i \in \Delta_k$ represent the flow of the i^{th} $\mathfrak{q}$ fragment of user request send through basestation k. The cost function for the i^{th} component of a user request can be expressed as follows:

$$w_k(i) = \mathfrak{q}\, S_k + i\, \mathfrak{q}^2. \quad (31)$$

In this case, the load balancing formulation can then be described as follows.

$$\min \sum_{k=1}^{C}\sum_{i=1}^{\Delta_k} w_k(i) y_{kK}^i \tag{32}$$

$$\text{s.t.} \sum_{k=1}^{C} \mathfrak{q} z_k^m = q_m, \ \forall m \in \mathcal{M} \tag{33}$$

$$\sum_{m=1}^{M} z_k^m - \sum_{i=1}^{\Delta_k} y_{kK}^i = 0 \ \forall k \in \mathcal{C} \tag{34}$$

$$-\sum_{k=1}^{C}\sum_{i=1}^{\Delta_k} \mathfrak{q} y_{kK}^i = -\sum_{m}^{M} q_m \tag{35}$$

$$z_k^m = \delta_k^m u_k^m, \ \forall k \in \mathcal{C}, m \in \mathcal{M} \tag{36}$$

$$\sum_{k=1}^{C} \delta_k^m = 1 \ \forall m \in \mathcal{M} \tag{37}$$

$$0 \le z_k^m \le u_k^m, 0 \le y_{kK}^i \le 1, \delta_k^m = \{0,1\} \tag{38}$$

The objective function (32) minimizes the load imbalance in the network. Constraints (33-35) are the flow conservation constraints. Constraints (36) ensure that the components of each request are serviced by a single BS. Therefore, the flow of the m^{th} user request is either 0 or u_k^m. Constraints (37) ensure that each request is serviced by a single BS. Equations (38) constraint the flow variables and define the binary indicator variables.

Note that in this case the binary variables δ_k^m are introduced in the formulation that significantly increase the computational complexity of the solution. However, as derived here, the problem formulation is valid for the general case where arbitrary user flows are requested and thus optimal forwarding decisions can be made.

Appendix C

Total unimodularity: proof of theorem 1: Let matrix C be the node-arc incidence matrix for the flow conservation constraints (24-26). Then the expressions $\{x, y \ge 0 : C[x; y] = b\}$ and $\{x, y \ge 0 : C[x; y] \le b, -C[x; y] \le -b\}$ are equivalent ways of describing these constraints. Further, let OUT and IN be the constraint matrices for the out-degree and in-degree constraints (equations (22) and (23), respectively), whose columns only correspond to links in L_1. For the capacity constraints, expressed by equation (27) in problem (P3), only links in L_1 and L_3 have unity values. Thus the capacity constraint matrix for links in L_1 and L_3 is the identity matrix, I. Initially we ignore the non-negativity constraints and consider them later in the proof. We define $A[x; y] \le z$ to be the matrix formed by the constraint equations (22)-(27), ignoring non-negativity and integrality constraints. Then A is characterized as follows:

$$A = \begin{array}{c} \overbrace{}^{L_1}\overbrace{}^{L_3}\overbrace{}^{L_2} \\ \begin{pmatrix} C & & \\ -C & & \\ I & & 0 \\ OUT & 0 & 0 \\ IN & 0 & 0 \end{pmatrix} \end{array} \tag{39}$$

We show that A is a *network matrix* and total unimodularity of network matrices is shown in [33]. A network matrix is defined as follows (further details can be found in [33]):

Definition S is a network matrix, if there exists a directed graph $Q = (V_q, E)$ and a directed spanning tree $T = (V_q, E(T))$ of Q, such that for each element $e = (v, w) \in E \setminus E(T)$ and $e' \in E(T)$, $S(e', e)$ is defined as follows:

$$S(e',e)=\begin{cases} +1 & v\text{-}w \text{ path in } T \text{ passes from } e' \text{ forwardly} \\ -1 & v\text{-}w \text{ path in } T \text{ passes from } e' \text{ backwardly} \\ 0 & v\text{-}w \text{ path in } T \text{ does not pass through } e' \end{cases} \tag{40}$$

The rows of S correspond to edges of the tree T and the columns of S correspond to non-tree edges of Q. Further, each column of S that corresponds to a non-tree arc $e = (v, w)$ traces the unique path from v to w on the tree T.

To show that A is a network matrix, a directed graph Q and a directed spanning tree T of Q are first generated. Tree T can be derived from the space-time network $G = (V, L)$ and the constraint matrix A by tracing the following steps:

1) Add initial node 0.
2) For every node in $i \in V$ of the space-time network add: nodes i, i' to T and arcs $e(i) = i \mapsto i'$ and $e(i') = 0 \mapsto i'$.
3) For every L_1 link $i \mapsto j$ of the space-time network:
   a) Add node $out(i)$ and arc $e(out(i)) = out(i) \mapsto i$, if not already available.
   b) Add node $in(j)$ and arc $e(in(i)) = j \mapsto in(j)$, if not already available.
4) For every L_1 link $i \mapsto j$ of the space-time network, add node $I(i, j)$ and arc $e(i, j) = I(i, j) \mapsto out(i)$.
5) For every L_3 link $m \mapsto n$ of the space-time network, add node $I(m, n)$ and arc $e(m, n) = I(m, n) \mapsto m$.

Note that the arcs of the tree T correspond to the rows of A. For the rows of matrix C, which are the nodes of the space-time network G, we add arcs $e(i) = i \mapsto i'$ and for the rows of $-C$ we add arcs $e(i') = 0 \mapsto i'$ (step 2). The rows of matrix I are the links $L_1 \cup L_3$ of G, and the rows of matrices OUT, IN are the nodes in V_p that have L_1 links incident to them. Corresponding to a link $i \mapsto j$ in L_1: for the rows in OUT we add arcs $e(out(i)) = out(i) \mapsto i$ to T, for the rows in IN we add arcs $e(in(i)) = j \mapsto in(j)$ (step 3), and for the rows in I we add arcs $e(i, j) = I(i, j) \mapsto out(i)$, $i \mapsto j \in L_1$ and $e(m, n) = I(m, n) \mapsto m$, $m \mapsto n \in L_3$ (steps 4 and 5). In this way, the rows of matrix A are mapped to arcs in tree T.

Further, the non-tree edges of Q are added. These edges correspond to the columns of matrix A, which are the arcs in L. To T, we add the following non-tree arcs as follows to construct graph Q:

1) For each link $i \mapsto j$ of G in L_1: add non-tree arc $I(i, j) \mapsto in(j)$ in Q.
2) For each link $i \mapsto j$ of G in L_2: add non-tree arc $i \mapsto j$ in Q.

3) For each link $m \mapsto n$ of G in L_3: add non-tree arc $I(m,n) \mapsto n$ in Q.

Having Q and T, we can check if A satisfies property (40) to be a qualified network matrix.

We demonstrate this by considering a few links of the space-time network G: $i \mapsto j, i \mapsto q \in L_1$, $k \mapsto l \in L_2$ and $m \mapsto n \in L_3$. The tree of this network is generated in figure 11 while the sub-matrices of A are shown in (41).

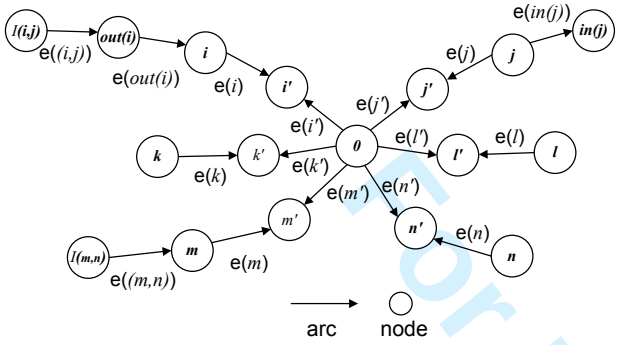

Fig. 11: Spanning tree T of directed graph Q for arcs $i \mapsto j$, $k \mapsto l$ and $m \mapsto n$ of G.

$$C = \begin{array}{c} \\ i \\ j \\ k \\ l \\ m \\ n \\ q \end{array} \begin{array}{c} \begin{array}{cccc} (i,j) & (i,q) & (k,l) & (m,n) \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 \end{bmatrix} \end{array} \quad I = \begin{array}{c} \\ (i,j) \\ (i,q) \\ (m,n) \end{array} \begin{array}{c} \begin{array}{ccc} (i,j) & (i,q) & (m,n) \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array} \tag{41a}$$

$$OUT = \begin{array}{c} \\ i \\ j \\ q \end{array} \begin{array}{c} \begin{array}{cc} (i,j) & (i,q) \end{array} \\ \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \end{array} \quad IN = \begin{array}{c} \\ i \\ j \\ q \end{array} \begin{array}{c} \begin{array}{cc} (i,j) & (i,q) \end{array} \\ \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array} \tag{41b}$$

The column of A that corresponds to an L_2 arc $k \mapsto l$ of G, corresponds to non-tree arc $k \mapsto l$ in Q. The unique path from k to l in T is as follows:

$$\begin{array}{ccccccccc} k & \to & k' & \to & 0 & \to & l' & \to & l \\ & +1 & & -1 & & +1 & & -1 & \end{array} \tag{42}$$

where the $+1$ and -1 indicate whether the arc is forward or backward in the path direction, respectively. Note that the column of A that corresponds to arc $k \mapsto l$ in Q will only have ± 1 at the exact rows that correspond to the tree arcs given by the path in (42). Further, we examine the column of A that corresponds to the L_1 arc $i \mapsto j$ of G. This column corresponds to non-tree arc $I(i,j) \mapsto in(j)$ in Q. The unique path from $I(i,j)$ to $in(j)$ in T can be traced as follows:

$$\begin{array}{ccccccccccc} I(i,j) & \to & out(i) & \to & i & \to & i\prime & \to & 0 & \to & j\prime & \to & j & \to & in(j) \\ & +1 & & +1 & & +1 & & -1 & & +1 & & -1 & & +1 & \end{array} \tag{43}$$

Lastly, a column in A with a link $m \mapsto n \in L_3$ of G, the non-tree arc is $I(m,n) \mapsto n$ in Q and the tree path can be traced as follows:

$$\begin{array}{ccccccccccc} I(m,n) & \to & m & \to & m\prime & \to & 0 & \to & n\prime & \to & n \\ & +1 & & +1 & & -1 & & +1 & & -1 & \end{array} \tag{44}$$

Once again, the corresponding non-zero values of the column in A traces the path generated in (44). Therefore, it is shown that this property holds for all links in L_1, L_2 and L_3 of graph G and thus A is a network matrix. As argued above, a network matrix is totally unimodular. Finally, A is TU if and only if $[A; I]$ is TU. Thus, including the non-negativity constraints preserves total unimodularity.

Panayiotis Kolios is a member of the Centre for Telecommunications Research at King's College London. His research interests revolve around the broad scope of wireless networking, Internet routing and multimedia communication. Applications of interest include mobile computing, transportation systems, emergency response and critical infrastructure protection. He is an active member of IEEE, contributing to a number of technical and professional activities within the Association. He received his B.Eng and Ph.D degrees in Telecommunications Engineering from King's College London, in 2008 and 2011 respectively.

Katerina Papadaki holds the position of Associate Professor of Management Science within the Department of Management at the London School of Economics. She received her Ph.D. in Princeton, Department of Operations Research and Financial Engineering. A major component of her research has been in developing algorithms to solve stochastic multidimensional dynamic programs and discrete combinatorial problems that arise in dynamic resource allocation problems with applications in wireless communication networks, transportation networks and finance. She is an active member of IEEE and INFORMS and she is associate editor of Optimization Letters.

Vasilis Friderikos lectures at the Centre for Telecommunications Research at King's College London in UK. His research interests lie broadly within the closely overlapped areas of wireless networking, mobile computing, and architectural aspects of the Future Internet. He has published more than 100 papers in the above scope research areas and actively involved in flagship IEEE conferences such as Globecom, ICC and PIMRC.